Algoritma optimisasi yang memenuhi differential privacy pada neural networks sebagai pencegahan serangan membership inference = Differentially private optimization algorithms for neural networks as the mitigation of membership inference attack / Roan Gylberth

Roan Gylberth, author

Abstrak

ABSTRAK
Neural networks merupakan salah satu pendekatan yang sering digunakan dalam melakukan analisis data. Dalam perkembangannya, neural networks mencapai kesuksesan dalam berbagai bidang, mulai dari pengenalan gambar, representasi bahasa,hingga bio informatika. Beberapa penelitian terakhir menunjukkan bahwa model neural networks memiliki kekurangan dalam melindungi informasi yang terdapat dalam training set agar tidak dapat dieksploitasi oleh pihak-pihak yang tidak berkepentingan. Kekurangan ini dapat dieksploitasi dengan membuat sebuah model yang dapat menentukan apakah seseorang berada dalam training set atau tidak, dan hasilnya dapat digunakan untuk melanggar privasi orang tersebut. Eksploitasi ini disebut dengan serangan membership inference. Serangan membership infrerence dapat dihindari oleh model yang memenuhi kriteria differential privacy, yaitu probabilitas keluaran dari model pada dua database yang berbeda pada satu baris pada dasarnya mirip. Pada tesis ini, dikembangkan algoritma optimisasi berbasis gradien seperti Momentum, Nesterov, RMSProp dan Adam yang memenuhi kriteria differential privacy. Algoritma yang dikembangkan digunakan untuk melatih model neural networks agar memenuhi kriteria differential privacy. Eksperimen yang dilakukan menunjukkan bahwa algoritma yang dikembangkan dapat digunakan untuk melatih model neural networks dan menghasilkan model yang lebih akurat dibandingkan algoritma stochastic gradient descent yang memenuhi kriteria differential privacy. Diperlihatkan juga pengaruh penjaminan privasi terhadap akurasi model yang dilatih menggunakan algoritma yang dikembangkan, yaitu penjaminan privasi yang lebih kuat menghasilkan akurasi model yang lebih rendah, dan sebaliknya.
<hr>
<i>ABSTRACT
Neural networks is one of the popular approach to analyze data. It has showed excellent ability to tackle complex problems in various domain, e.g., computer vision,language representation, and bioinformatics. At some point, neural network model may leak some information about the training data. This leakage could be exploited by adversaries to violate individuals in the training data. Membership inference attack is one kind of attacks that could be used by the adversary. This attack can be mitigated by using differentially private models. In this thesis, differentially private optimization algorithms, i.e., momentum, nesterov, rmsprop, adam, were developed. These algorithms then used to train a differentially private neural networks model. It was shown by the experiments conducted that these algorithms can be used to train a neural networks model, and yields better model accuracy compared to stochastic gradient descent algorithm. The tradeoff between privacy and utility is also studied.